Semantically Aware Text Categorisation for Metadata Annotation

Semantically Aware Text Categorisation for Metadata Annotation

Giulio Carducci*†, Marco Leontino◇, Daniele P. Radicioni◇, Guido Bonino*, Enrico Pasini*, and Paolo Tripodi*

◇Dipartimento di Informatica, Università degli Studi di Torino
*Dipartimento di Filosofia, Università degli Studi di Torino
†giulio.carducci@protonmail.com,
{name.surname}@unito.it

Abstract. In this paper we illustrate a system aimed at solving a long-standing and challenging problem: acquiring a classifier to automatically annotate bibliographic records by starting from a huge set of unbalanced and unlabelled data. We illustrate the main features of the dataset, the learning algorithm adopted, and how it was used to discriminate philosophical documents from documents of other disciplines. One strength of our approach lies in the novel combination of a standard learning approach with a semantic one: the results of the acquired classifier are improved by accessing a semantic network containing conceptual information. We illustrate the experimentation by describing the construction rationale of training and test set, we report and discuss the obtained results and conclude by drawing future work.

Keywords: Text Categorization, Lexical Resources, Semantics, NLP, Language Models

1 Introduction

To date natural language processing (NLP) resources and techniques are being used in many tasks, such as conversational agents applications [22], question answering [13], automatic summarization [15], keywords extraction [24], text categorisation [17]. In this paper we propose a system to automatically annotate metadata related to scholarly records; in particular, we show how lexical resources can be paired to standard categorisation algorithms to obtain accurate categorisation results.

This work is carried out in the frame of a broader philosophical research project aimed at investigating a set of UK doctoral theses collected by the Electronic Theses Online Service (EThOS).[1] Although we presently consider only the EThOS dataset, a huge amount of such documents have been collected within the project activities from different sources and countries, such as US, Canada, Italy, and other PhD theses are currently being searched to collect further data.

[1] https://ethos.bl.uk.

Of course, many issues arise when trying to apply a uniform data model to such heterogeneous data; data is noisy, with partly missing information (e.g., abstracts are mostly missing until more recent years), and so on. Amongst the most basic issues, we single out a problem of text categorisation. In fact, when searching for philosophical theses in the EThOS dataset (i.e., those with 'Philosophy' in the `dc:subject` field) not all retrieved records are actually related to Philosophy, but rather to cognate disciplines such as Sociology, Religion, Psychology, and so forth. Additionally, in some cases the subject field is empty, or it contains numbers, or different sorts of noisy information. The thesis subject may be of little relevance in this setting because in UK there is no clear and univocal administrative classification of PhD titles according to disciplines. We presently focus on the problem of categorising such records in order to further refine the information provided by the `dc:subject` field, and to individuate philosophical theses. Although the task at hand is a binary classification problem, it is not that simple in that *i)* in many cases the thesis disciplines are distinct though not well separated; *ii)* the abstract may be lacking (thus very little information is available), and *iii)* no labelled data is available to train some learning algorithm.

Although the methodology described in the paper has been developed to cope with a specific problem, the proposed solution is general; the whole system basically implements an attempt at integrating domain specific knowledge (acquired by training a learning system) and general knowledge (embodied in the BabelNet semantic network). Specifically, we show that the output of a state-of-the-art algorithm (Random Forest [14], an ensemble learning technique building on decision trees) trained on a specific dataset can be refined through a search over a semantic network grasping general conceptual knowledge. The obtained results significantly improve on those provided by the two software modules separately. Also, the system enjoys the nice property of providing a concise explanation illustrating why a thesis should be considered as properly philosophical, based uniquely on the information available in the thesis title.

The paper is structured as follows: in Section 2 we briefly survey the related work on text categorisation; we then describe the EThOS dataset and provide some descriptive statistics to qualify it (Section 3). The System is then illustrated in full detail (Section 4). In Section 5 we present and discuss the results of the evaluation of the system —which was tested on a dataset handcrafted by two human experts— and conclude by pointing out present weaknesses and future work (Section 6).

2 Related Work

Classification of textual documents is a task that draws particular interest in the field of natural language processing and is characterised by numerous challenges such as the high dimensionality and sparsity of the data, unbalanced classes, the lack of enough annotated samples and the time and effort required to manually inspect large datasets.

During the years, many techniques have been proposed that address one or more of the mentioned challenges trying to reduce their negative impact. Traditional approaches usually feature machine learning algorithms such as decision trees, random forests [3, 8], support vector machines (SVM) [16, 2], Naïve Bayes [25, 5]. Some hybrid approaches have been proposed that combine the result of the classification with other types of information. Wang integrates the classification of documents and the knowledge acquired from Chinese digital archives into a concept network, enhancing the metadata of documents and their organisation in digital libraries [32]. Similarly, Ferilli *et al.* build a semantic network from the text, where concepts are connected by a set of relationships, consisting in verbs from the input documents [9]. The resulting taxonomy can be used to perform some semantic reasoning or understand the content of the documents, although it needs some refinement. Nigam *et al.* address the problem of scarcity of annotated data by combining the Expectation-Maximization (EM) and a Naïve Bayes classifier [29]. They show how the classification error can be significantly reduced by an appropriate weighing of the documents and modelling of the classes. Gabrilovich *et al.* propose a classifier that is able to match documents with Wikipedia articles, then integrate the concepts extracted from such articles into the features of the original document, thus enriching its semantic representation [11].

Textual data is often represented using bag-of-words (BoW) techniques, in which a given document is mapped onto a high-dimensional vector space [30]. The resulting vector can then be used to train a linear classifier, for example Linear Regression or SVM. There is also a significant volume of research in literature on the use of Random Forests; Cutler *et al.* show that Random Forests outperform linear methods for three ecology-related tasks [7]; Akinyelu and Adewumi achieve an high accuracy on classification of phishing emails using a combination of meticulously defined features [1] , while Xu *et al.* optimise the classification accuracy by weighing the input features of the individual trees and excluding the ones that negatively affect the performance of the classifier [34].

A rather novel language modeling technique consists in word embeddings, that is, dense vector representation of words, that have the property of carrying semantic information about words and their context. Word embeddings gained increasing interest in the latest years and are normally employed in a deep learning context as in [18, 19]: documents are transformed using a pretrained dictionary and are processed by the neural network which ultimately outputs a class label.

3 Dataset and Gold Standard

The EThOS initiative is aimed at sharing information about UK Doctoral theses and at "making the full texts openly available for researchers".[2] The dataset used in this work consists of a *corpus* of PhD theses whose publication dates range

[2] `http://ethostoolkit.cranfield.ac.uk`.

Table 1: Statistics about the textual content of the dataset. The values reported are computed on subject, title, and (if present) abstract of the record.

Measure	**With Abstract**	**Without abstract**	**Whole dataset**
Words	64,946,071	3,480,858	68,426,929
Words after preprocessing	37,875,285	2,548,988	40,424,273
Unique words	1,496,488	258,045	1,610,896
Unique words after preprocessing	435,825	124,650	476,769
Average words per record	321.96	12.72	143.94
Average words per record after preprocessing	187.76	9.31	85.03

from the second half of the Twentieth Century to the most recent years. Such *corpus* has been kindly made available by the staff of the EThOS service of the British Library, and consists in nearly half a million bibliographic records (namely 475,383); records with empty abstract are 57.6% (overall 273,665), while the abstract is present in 42.4% of such theses (that is, 201,718). The corpus implements the following metadating schema:

- `uketdterms:ethosid`: the identifier of the record within the EThOS digital library;
- `dc:title`: the title of the thesis;
- `dc:creator`: the PhD student who authored the thesis;
- `uketdterms:institution`: the name of the University;
- `dc:publisher`: may differ from institution in some cases;
- `dcterms:issued`: year of publication;
- `dcterms:abstract`: abstract of the thesis (when available);
- `dc:type`: always "Thesis or Dissertation";
- `uketdterms:qualificationname`: "Thesis", sometimes followed by area of study and University;
- `uketdterms:qualificationlevel`: "Thesis", "Doctoral" or similar;
- `dc:identifier`: pointer to the resourse in EThOS digital library;
- `dc:source`: pointer to the original location of the resourse (e.g., institutional website);
- `dc:subjectxsi`: empty for all records;
- `dc:subject`: synthetic description of the area of study.

The above schema employs three different vocabularies to define the metadata of a document. Dublin Core Metadata Element Set (`dc`) and its extension DCMI Metadata Terms (`dcterms`) are both defined in the Dublin Core Schema [33], which features a number of attributes that can be used to describe a digital or physical resource within a collection (e.g., a book in a library), while the uketd_dc namespace (`uketdterms`) is defined on top of `dcterms` and describes the core set of metadata for UK theses that are part of the EThOS dataset.[3]

[3] Full account of the EThOS UKETD_DC application profile can be found at the URL `http://ethostoolkit.cranfield.ac.uk/tiki-index.php?page=Metadata`.

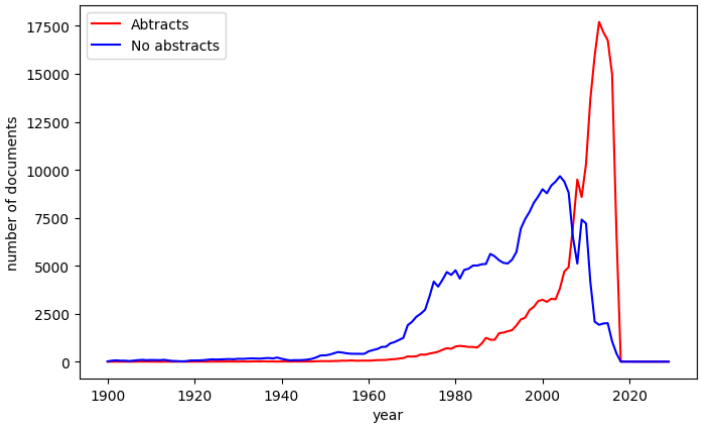

Fig. 1: Distribution of the number of theses by year of publication in the dataset. The corpus with abstracts counts less elements and they are distributed in a shorter and more recent time period.

Some descriptive statistics about the textual content of the records are reported in Table 1 that details total words, total unique words and average number of words per element, before and after preprocessing of the text. Such figures are computed also on the two subsets individuated based on the distinction between empty/valued abstract. We considered three fields: `dc:subject`, `dc:title` and `dcterms:abstract`, when available. We note that the presence *vs.* absence of the abstract is a key aspect for a record; in fact, records containing abstract information account for almost 95% of the total word count.

We note that the preprocessing phase has higher impact on the elements containing abstract information (where we observe 42% reduction of the available words after the preprocessing step) with respect to the second one (27% decrease). This is because titles and subjects tend to be shorter and more synthetic, sometimes consisting only in a list of keywords or concepts, in contrast with abstracts that are more exhaustive and written in fully fledged natural language. This tendency is in fact evident in Table 1; the average number of words increases by a factor of 25 when the abstract is present. Finally, in Figure 1 we plotted the distribution of the number of theses per year of publication. The graph is computed separately for the two subsets of data (with-without abstract information), and clearly shows that the records with abstract are more recent and concentrated in a smaller time span. Both distributions have their peak after year 2000, and drop right before 2020, in accord with intuition.

4 The System

The goal of our work is to identify as many philosophy theses as possible: as mentioned, the problem at hand is that of individuating the theses that are

Table 2: Statistics of the dataset, partitioned into philosophical and non-philosophical records based on the educated guess.

Corpus	**Philosophy**	**Not Philosophy**	**Total**
With abstracts	1,495	200,223	201,718
Without abstracts	1,982	271,683	273,665
Whole dataset	3,477	471,906	475,383

actually related to philosophy, by discarding all records related to similar though distinct disciplines. The original dataset is not annotated and so it does not contain any explicit information that we could use to pursue our goal. We had two options: *i)* to apply an unsupervised learning technique on the structured data, such as clustering or topic modelling, in order to single out philosophy samples among the multitude of subjects; or *ii)* to automatically annotate the data (i.e., an educated guess), and subsequently to use such annotated data to train a supervised learning algorithm. We chose the second option, and set up a binary classification framework where a record from the dataset is classified as either philosophical or non-philosophical. We hypothesised that a supervised model would have fit our use case way better that an unsupervised one. In fact, we can train the algorithm by providing specific examples of the two classes, which will result in a better-defined decision boundary.

In the following we first describe the data employed and how we built a training set (based on a educated guess) and a test set (annotated by human experts). We then illustrate the training of a binary classifier (hereafter Random Forest Module), and elaborate on the learnt features. Finally, we introduce the Semantic Module, devised to refine the predictions of the Random Forest Module.

4.1 Building the training set and the test set

The first step was devised to build the training and the test set based on an educated guess. This bootstrapping technique is a key aspect of the whole approach, as it allowed us to adopt supervised machine learning algorithms. In many settings, in fact, a first raw, automatic categorisation can be attempted in order to overcome the limitation due to the lack of labelled data. Given the huge number of documents in the *corpus*, we did not consider the option of manually annotating the data in order to select a significant sample. We instead employed a text-based extraction method to search for relevant documents in the dataset by using a regular expression. We searched in the field `dc:subject` of each document and selected all samples matching the keyword *philosophy*, or a meaningful part of it (i.e., the substring *philosop*). We consider such documents as positive examples, while we consider all the other ones as negative examples. Of course this strategy is not completely fault-free, in fact it is possible that a given document is philosophical even though there is no *philosophy* specified in the subject (e.g., *Kant's reasoning*).

Table 3: Basic statistics about the textual content of the training set. The values reported are computed only on title.

Measure	Value
Total words	371,478
Total words after preprocessing	249,266
Total unique words	53,338
Total unique words after preprocessing	34,549
Average words per document	11.41
Average words per document after preprocessing	7.65

Table 2 illustrates how the records in the dataset have been classified, based on this simple partitioning rule. We observe that only $3,477$ records (that is 0.73% of the whole dataset), were initially recognised as pertaining to philosophy theses, while the remaining $471,906$ are negative examples from all other fields of study.

Building the training set We then left out 500 randomly chosen records from the positive examples and as many records from the negative examples that were used at a later time in order to build a test set. The final training set is thus composed of $2,977$ positive and $471,406$ negative examples. Not all negative examples were actually used to train the classifier: the training set has been built by randomly selecting a number of negative samples that outnumbers its positive counterpart by a factor of 10, thereby resulting in $29,770$ records. Some descriptive statistics of the training set are reported in Table 3.

Building the test set To create the test set we randomly selected 500 documents from each of the two groups, thus creating a new set of $1,000$ samples. Such samples were manually annotated by two domain experts. Interestingly enough, even though they had access to the whole record (different from our system, that only sported the `dc:title` field), in some cases (around 20 out of thousand records) the domain experts could not make a clear decision. All such ambiguous cases were left out from the test set. We did not record the inter-annotator agreement, since only the records where the annotators agreed were retained.[4] The final test set was built by adding further randomly chosen records that were annotated by the experts, finally obtaining a balanced set of 500 philosophical records and 500 non-philosophical records.

4.2 The Random Forest Module

We then trained the classifier to acquire a model for the categorisation problem at hand.[5] In order to train the classifier we considered only the terms in the

[4] The final test set is available within the bundle containing the implementation of the described system [4].

[5] An off-the shelf implementation of the Random Forest algorithm was used, as provided by the scikit-learn framework, `http://scikit-learn.org/stable/`.

`dc:title` field, which is available in all records of the dataset and that in almost all cases suffices also to human experts to classify the record. A preprocessing step was devised, in order to filter out stopwords and to normalise the text elements (please refer to Table 2 reporting the statistics of the dataset after the preprocessing step). We chose not to use abstracts to train the model, since they are not available for most records; nor we used such information at testing time, even if available. By doing so, we adopted a conservative stance, and we are aware that some helpful information is not used, thereby resulting in a lower bound to the performance of the categorisation system. The applied preprocessing steps are:

1. **Conversion to lowercase** "Today I baked 3 apple pies!" → "today i baked 3 apple pies!".
2. **Stop-words removal**[6] "today i baked 3 apple pies!" → "today baked 3 apple pies!".
3. **Punctuation removal** "today baked 3 apple pies!" → "today baked 3 apple pies".
4. **Numbers removal** "today baked 3 apple pies" → "today baked apple pies".
5. **Stemming**[7] "today baked apple pies" → "today bake apple pie".
6. **Short document removal** documents with $n_{tokens} < 3$ are removed.

After preprocessing, we transformed the documents into vectors using a bag-of-words approach that maps each of them to the vector space with the tf-idf transformation. Tf-idf computes the frequency of terms in a document, weighed by the number of documents (within a given collection) that contain such term. This procedure favours important and more discriminative terms rather than common ones. The resulting dictionary vector (containing all meaningful terms in the collection) has been truncated to $60,000$ features, based on the frequency of the terms in the corpus, so to discard highly uncommon ones.

The estimator that we employed to acquire a binary classifier is Random Forest [14]. This is an ensemble method that trains a set of decision trees by using random subsets of input features, and assigns to a given sample the class that is predicted more often among the different classifiers. This choice is motivated by the fact that Random Forest can handle a large number of features and provides a measure of their importance; this may be of particular interest to examine the intermediate stages of the computation. The output of the classifier includes the set of terms that are most probably predictive for a sample to be positive or negative for a given class label. However, several algorithms could be used in principle in this step, by plugging a different learner into the overall system.

We preprocessed each document in the training set and extracted the corresponding vector representation along with its label (which was computed based on the educated guess, as illustrated above). Given the binary categorisation

[6] We used the list of English stop-words from the NLTK package available at the URL `https://gist.github.com/sebleier/554280`.

[7] Stemming was done using the WordNet Lemmatizer, also available within the NLTK library, `https://github.com/nltk/nltk/blob/develop/nltk/stem/wordnet.py`.

Table 4: Configuration parameters used for the Random Forest classifier. Namely, 50 decision trees were trained, each of them assigning either a class label 0 or 1 to a given vector. The final class label will be the most frequent one. Other parameters are kept as their default value.

Parameter	**Value**
Number of estimators	50
Max features	0.6
Random state	none
Max depth	none

setting, labels did encode only two classes: 'philosophy' and 'non-philosophy'. The training set was fed to the estimator, to extract significant patterns in the data and to learn how to exploit them to individuate philosophy theses. Table 4 shows some relevant configuration parameters.

Figure 2 reports the 30 most important features, along with a relevance score, ranging over $[0, 1]$. The score of a feature is computed for a single tree of the forest as the total decrease of node impurity brought by that feature, and is averaged among all trees. We also report the standard deviation of their values. We observe that the majority of such terms is highly predictive of a philosophical context, even though among the most relevant learnt features also terms proper to the Religion class are present (e.g., 'theology', 'church', 'religious', 'biblical').[8]

For this reason we further investigated the score acquired for the features, with particular focus to philosophy-related ones:[9] in Figure 3 we show 30 salient philosophical terms, whose score is lower than that learnt for the term 'biblical', which is the rightmost feature portrayed in Figure 2. Such scores were likely to negatively affect the recall of the Random Forest Module, which acquired inaccurate weights, probably due to the scarcity of philosophical training data and to the noise present in the educated guess. This is why we devised the other module that —independent of the information in the training set— relies on the knowledge available in BabelNet, as described in the following.

4.3 The Semantic Module

The semantic module performs some basic Information Extraction tasks, accessing the lexical conceptual resource of BabelNet [28]. BabelNet is a multilingual semantic network resulting from the integration of WordNet and Wikipedia; it builds on the constructive rationale of WordNet —that is, it relies on sets of synonyms, the Babel synsets— which is extended through the encyclopedic structure

[8] It is worth noting that the human experts adopted a rather inclusive attitude with respect to religious studies, based on their previous acquaintance with an analogous dataset of US PhD dissertations, in which a significant number of 'religious' dissertations have been defended in philosophy departments.

[9] We obtained a list of some relevant philosophical concepts from the upper levels of the Taxonomy of Philosophy by David Chalmers, `http://consc.net/taxonomy.html`.

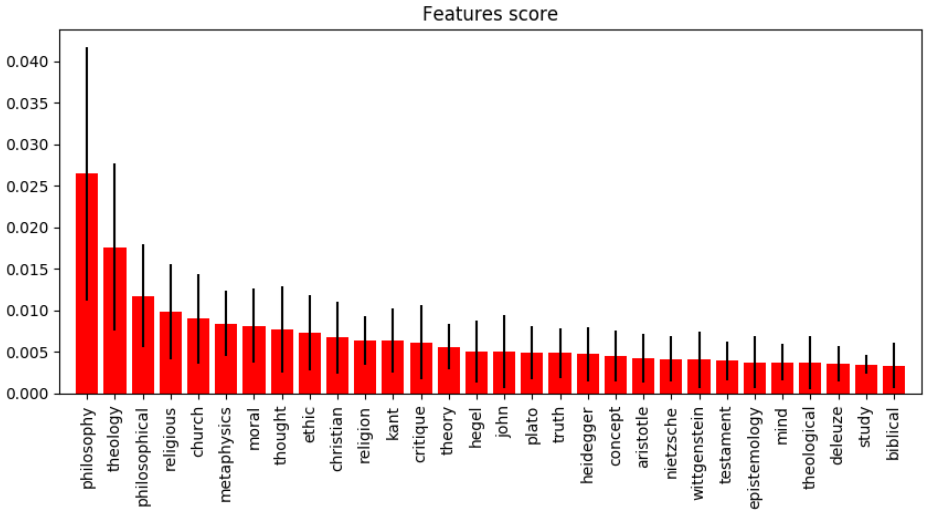

Fig. 2: The top 30 most discriminating features of the classifier: red bars report the average relevance score among the trees, and black bars report the standard deviation.

of Wikipedia. In particular, the nodes in the network represent concepts and entities (that is, persons, organisations and locations), and the edges intervening between each two nodes represent semantic relations (such as *IsA*, *PartOf*, *etc.*). Although further lexical resource exist containing different sorts of knowledge (such as, e.g., WordNet [27], ConceptNet [23], COVER [20, 26], or a hybrid approach proposed by [12, 21]), we chose to adopt BabelNet in that it ensures a broad coverage to concepts and entities as well, that in the present domain are particularly relevant. The semantic module aims at searching the terms present in the theses title, to individuate the underlying concept and then at checking whether they are either philosophical concepts (that is, linked to 'philosophy' in the BabelNet taxonomy) or philosophers. It performs three steps: named entities recognition (NER), multiwords expressions (MWEs) extraction, and BabelNet search, that are illustrated in the following.

NER. At first, named entities are extracted;[10] out of all recognised entities, only persons are retained.

MWEs extraction. We first perform Part-Of-Speech (POS) tagging[11] of each record title, and we then select both individual NOUNs and multi-word expressions. MWEs are extracted based on few patterns: NOUN+NOUN (e.g., matching expressions such as 'belief revision', or 'quantum theory'); ADJ+NOUN (e.g., matching 'analytic philosophy'); and NOUN+PP+NOUN

[10] We presently employ the Stanford Named Entity Recognizer [10].
[11] We presently employ the Stanford POS Tagger [31].

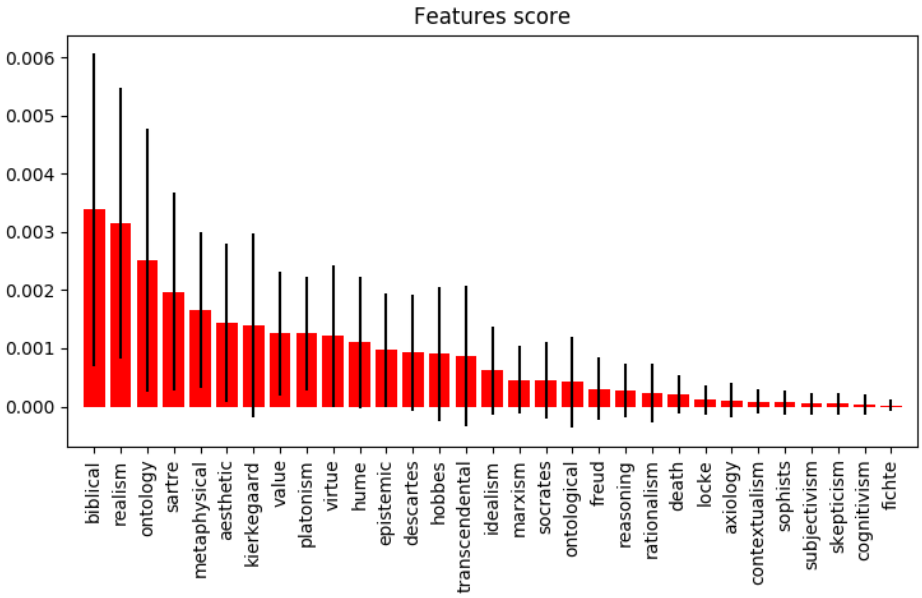

Fig. 3: Score of 30 additional philosophy-related terms compared to the word 'biblical'.

('philosophy of mind'). Such lexical elements are enriched with their conceptual counterpart in the subsequent step.

BabelNet search. The previously extracted terms are then searched in BabelNet, and their corresponding synsets (that is, sets of synonyms along with their meaning identifiers) retrieved. At this stage of the computation, we discard the MWEs that are not present in BabelNet, thus implementing a semantic filtering for the terms individuated through the patterns described above. For each sense_t or entity_t associated to each extracted term t we inspect if it corresponds to philosophy (bn:00061984n) or philosopher (bn:00061979n), and whether it is linked to either concept. In doing so, we basically explore the relations *IsA* and *Occupation*, and we retain any sense_t and entity_t such that

- [sense_t,entity_t] *IsA* philosophy/philosopher; or
- [sense_t,entity_t] *Occupation* philosopher.

Building the PHILO-ENTITIES array. Such elements are added to the description of the record for the thesis being classified, in the philosophical-entities set. We note that thanks to the linking of BabelNet synsets with external triple stores (such as Wikidata), such triples can be exploited to perform further analysis of the record, and of the entities herein contained.

Table 5: Categorisation results obtained on the test set by experimenting with the Random Forest Module, with the Semantic Module, and with their combination.

	RF	**SEM**	**RF+SEM**
Precision	**0.8227**	0.8269	0.7944
Recall	0.5660	0.6400	**0.7880**
F1	0.6706	0.7215	**0.7912**
Accuracy	0.7220	0.7530	**0.7920**

The decision rule of the semantic module is very simple: if the array of philosophical entities associated to this record is not empty, we label it as a philosophical one; we label it as a non-philosophical one, otherwise.

The set of PHILO-ENTITIES can be used to build simple yet informative explanations of why a given record has been categorised as a philosophical one. This approach, based on simple templates such as the system described in [6] will be extended to build explanations also for non-philosophical records in next future. Let us consider, as an example, a record whose title is "Dialectic in the philosophy of Ernst Bloch"; this record has been marked as philosophical by human annotators. While processing this record, the Semantic Module detects the concepts philosophy (bn:00061984n) and dialectic (bn:00026827n) as associated to the concept philosophy, as well as the Named Entity Ernst Bloch (bn:03382194n) as a person whose occupation is that of philosopher.

The semantic module is executed when the first module (implementing the Random Forest-based classifier) returns 0, that is when the record is not recognised as a philosophical one in the first stage of the computation.

5 Evaluation

We evaluated the system on a test set composed of 1,000 records, annotated by human experts, and built as described in Section 4.1. All modules of our system were run with only title information in input[12]. In the experimentation we recorded *i*) the results of the Random Forest Module alone (which is a state-of-the-art algorithm, thus working as our baseline); *ii*) the results of the Semantic Module alone; and *iii*) the results of both modules, where the latter module is executed only in case a record is predicted to be non-philosophical by the former one. The results are presented in Table 5.

Discussion The system obtained encouraging results. First of all, as earlier mentioned, the dataset was strongly unbalanced, with a vast majority of records that were non-philosophical (please refer to Table 2), but with many records coming from closely related research fields. Namely, out of the overall 475K records, those concerned with philosophy were less than 3.5K, thus in the order of 0.7%. Yet, to conduct a thorough experimentation we restricted to considering only

[12] The implemented system is delivered through the Zenodo platform [4].

title information, thus often exploiting only a fraction of the available information. To consider in how far this limitation can be harmful to categorisation, let us consider that the human experts in some cases were not able to decide (or to decide consistently) the class of the considered records: when available, they had the opportunity to inspect the abstract and any other field of the record. Additionally, the assumption underlying the Semantic Module was rather fragile: just looking for people's occupation and for concepts hypernyms is a crude way to determine whether philosophy (or any other discipline) is mentioned. It was necessary to avoid more noisy relations in BabelNet (such as *SemanticallyRelated*), that allow retrieving many more entities connected to the concept at hand, but in less controlled fashion.

We observe that the Random Forest Module obtains a high precision, at the expense of a poor recall, caused by a significant number of false negatives, as expected by inspecting the feature weights. The attempt at correcting this behaviour has been at the base of the design of the semantic module; specifically, we strove to reduce the number of false negatives, meantime limiting the growth of the false positives. The key of the improvement in the recall (over 22%) obtained by the whole system with respect to the Random Forest Module is thus easily explained: the whole system incurs in false negatives in less than half cases, with a reduced increase of false positives.

Provided that the explanatory features of the system were not object of the present experimentation, nonetheless we briefly report on this point, too, as about a preliminary test. An explanation is built only when a record is associated to either some philosopher(s) or to philosophical concept(s): it is thus presently conceived simply as a listing of the elements collected in the PHILO-ENTITIES array. The system generated overall 496 explanations: in 394 cases this correctly happened for a philosophical record (thus in 78.8% of cases), whilst in 102 cases an explanation was wrongly built for non-philosophical records. Interestingly enough, when both modules agree on recognising a record as a philosophical one, the PHILO-ENTITIES array contains on average 1.78 elements; when the Random Forest Module predicts 'non-philosophical' label and Semantic Module (correctly) overwrites this prediction, the PHILO-ENTITIES array contains on average 1.57 elements. Thus less information is available also to the Semantic Module, which can be interpreted as a recognition that records misclassified by the Random Forest Module are objectively more difficult. However, further investigation is required to properly interpret this datum, and to select further semantic relations in BabelNet.

6 Conclusions

In this paper we have presented a system for categorising bibliographic records, to automatically characterise the metadata about the subject of the record. The research question underlying this work was basically how to integrate domain specific knowledge (acquired by training a learning system) and general knowledge (embodied in the BabelNet semantic network). As we pointed out,

the difficulty of the present task was caused by the unfavourable bootstrapping conditions.

We have described the EThOS dataset, and illustrated the methodology adopted: based on the educated guess, we tentatively classified all records. After partitioning the data between training and test set, we trained a Random Forest learner to acquire a classifier for the training set. On the other side, we developed the Semantic Module, which is charged to extract concepts and entities from the title field of the records, exploiting the BabelNet semantic network. The evaluation revealed that the system integrating both modules works better than the individual software modules: we obtained interesting results. Future work will include improving the explanation, exploring additional semantic relations, and considering further knowledge bases.

Acknowledgments

The authors wish to thank the EThOS staff for their prompt and kind support. Giulio Carducci and Marco Leontino have been supported by the REPOSUM project, BONG_CRT_17_01 funded by Fondazione CRT.

References

1. Akinyelu, A., Adewumi, A.: Classification of phishing email using random forest machine learning technique 2014 (04 2014)
2. Begum, N., Fattah, M., Ren, F.: Automatic text summarization using support vector machine 5, 1987–1996 (07 2009)
3. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (Oct 2001)
4. Carducci, G., Leontino, M., Radicioni, D.P.: Semantically Aware Text Categorisation for Metadata Annotation: Implementation and Test Files (2018), `https://doi.org/10.5281/zenodo.1446128`
5. Chen, J., Huang, H., Tian, S., Qu, Y.: Feature selection for text classification with naïve bayes. Expert Syst. Appl. 36(3), 5432–5435 (Apr 2009)
6. Colla, D., Mensa, E., Radicioni, D.P., Lieto, A.: Tell me why: Computational explanation of conceptual similarity judgments. In: et al., J.M. (ed.) Proceedings of the 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Special Session on Advances on Explainable Artificial Intelligence. Communications in Computer and Information Science (CCIS), vol. 853, pp. 74–85. Springer International Publishing, Cham (2018)
7. Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J.: Random forests for classification in ecology. Ecology 88 11, 2783–92 (2007)
8. Fatih Amasyali, M., Diri, B.: Automatic turkish text categorization in terms of author, genre and gender (05 2006)
9. Ferilli, S., Leuzzi, F., Rotella, F.: Cooperating techniques for extracting conceptual taxonomies from text (01 2011)

10. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd annual meeting on association for computational linguistics. pp. 363–370. Association for Computational Linguistics (2005)
11. Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In: Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2. pp. 1301–1306. AAAI'06, AAAI Press (2006)
12. Ghignone, L., Lieto, A., Radicioni, D.P.: Typicality-Based Inference by Plugging Conceptual Spaces Into Ontologies. In: Procs. of AIC. CEUR (2013)
13. Harabagiu, S., Moldovan, D.: Question answering. In: The Oxford Handbook of Computational Linguistics. Oxford University Press (2003)
14. Ho, T.K.: Random decision forests. In: Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1. pp. 278–. ICDAR '95, IEEE Computer Society, Washington, DC, USA (1995)
15. Hovy, E.: Text summarization. In: The Oxford Handbook of Computational Linguistics 2nd edition. Oxford University Press (2003)
16. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) Machine Learning: ECML-98. pp. 137–142. Springer Berlin Heidelberg, Berlin, Heidelberg (1998)
17. Johnson, R., Zhang, T.: Semi-supervised convolutional neural networks for text categorization via region embedding. In: Advances in neural information processing systems. pp. 919–927 (2015)
18. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics (2017)
19. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. pp. 2267–2273. AAAI'15, AAAI Press (2015)
20. Lieto, A., Mensa, E., Radicioni, D.P.: A Resource-Driven Approach for Anchoring Linguistic Resources to Conceptual Spaces. In: Procs. of the 15th Conference of the Italian Association for Artificial Intelligence. LNAI (10037), vol. 10037, pp. 435–449. Springer (2016)
21. Lieto, A., Radicioni, D.P., Rho, V.: Dual peccs: a cognitive system for conceptual representation and categorization. Journal of Experimental & Theoretical Artificial Intelligence 29(2), 433–452 (2017), http://dx.doi.org/10.1080/0952813X.2016.1198934
22. Lison, P., Kennington, C.: Opendial: A toolkit for developing spoken dialogue systems with probabilistic rules. Proceedings of ACL-2016 System Demonstrations pp. 67–72 (2016)
23. Liu, H., Singh, P.: Conceptnet-a practical commonsense reasoning tool-kit. BT technology journal 22(4), 211–226 (2004)
24. Marujo, L., Ribeiro, R., de Matos, D.M., Neto, J.P., Gershman, A., Carbonell, J.: Key phrase extraction of lightly filtered broadcast news. In: Proceedings of 15^{th} International Conference on Text, Speech and Dialogue (TSD 2012). Springer (September 2012)
25. McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. In: IN AAAI-98 Workshop on Learning for Text Categorization. pp. 41–48. AAAI Press (1998)

26. Mensa, E., Radicioni, D.P., Lieto, A.: COVER: a linguistic resource combining common sense and lexicographic information. LANG RESOUR EVAL (Jun 2018)
27. Miller, G.A.: WordNet: a lexical database for english. Communications of the ACM 38(11), 39–41 (1995)
28. Navigli, R., Ponzetto, S.P.: BabelNet: Building a very large multilingual semantic network. In: Procs. of the 48th ACL. pp. 216–225. ACL (2010)
29. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using em. Mach. Learn. 39(2-3), 103–134 (May 2000)
30. S. Harris, Z.: Distributional structure 10, 146–162 (08 1954)
31. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13. pp. 63–70. Association for Computational Linguistics (2000)
32. Wang, J.: A knowledge network constructed by integrating classification, thesaurus, and metadata in digital library. International Information & Library Review 35(2-4), 383–397 (2003)
33. Weibel, S.: The dublin core: a simple content description model for electronic resources. Bulletin of the American Society for Information Science and Technology 24(1), 9–11 (1997)
34. Xu, B., Guo, X., Ye, Y., Cheng, J.: An improved random forest classifier for text categorization 7 (12 2012)